Nucleic Acids Research, 2018 ***1–14***
doi: 10.1093/nar/gkx1321

PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures

Pawel S. Krawczyk[1,2,*], Leszek Lipinski[1] and Andrzej Dziembowski[1,2]

[1]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warsaw, Poland and [2]Department of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Pawinskiego 5a, 02-106 Warsaw, Poland

Received June 12, 2017; Revised December 01, 2017; Editorial Decision December 22, 2017; Accepted December 28, 2017

ABSTRACT

Plasmids are mobile genetics elements that play an important role in the environmental adaptation of microorganisms. Although plasmids are usually analyzed in cultured microorganisms, there is a need for methods that allow for the analysis of pools of plasmids (plasmidomes) in environmental samples. To that end, several molecular biology and bioinformatics methods have been developed; however, they are limited to environments with low diversity and cannot recover large plasmids. Here, we present PlasFlow, a novel tool based on genomic signatures that employs a neural network approach for identification of bacterial plasmid sequences in environmental samples. PlasFlow can recover plasmid sequences from assembled metagenomes without any prior knowledge of the taxonomical or functional composition of samples with an accuracy up to 96%. It can also recover sequences of both circular and linear plasmids and can perform initial taxonomical classification of sequences. Compared to other currently available tools, PlasFlow demonstrated significantly better performance on test datasets. Analysis of two samples from heavy metal-contaminated microbial mats revealed that plasmids may constitute an important fraction of their metagenomes and carry genes involved in heavy-metal homeostasis, proving the pivotal role of plasmids in microorganism adaptation to environmental conditions.

INTRODUCTION

Plasmids are mobile genetic elements that facilitate rapid evolution and adaptation of their hosts under changing environmental conditions (1,2). Plasmids are extra-chromosomal fragments of DNA that replicate autonomously in the host cell and are widely represented in bacterial species. Most of the known plasmids occur in circular form, which is an important feature allowing for their easy isolation using the alkaline lysis method. There are, however, also representatives, mostly from the *Borellia*, *Streptomyces, Nocardia* and *Rhodococcus* genera, which are linear (3). An important feature of plasmids is that they usually lack genes commonly assigned to primary metabolic processes but rather possess genes improving environmental fitness of the host or coding catabolic or resistance functions (4–6). Moreover, they can contribute to horizontal gene transfer between different species from diverse taxonomic groups, which make them factors with significant ecological impact (7). Therefore, plasmid-oriented studies are important to better understand processes occurring in diverse environments, and there is a need for methods to identify and sequence new plasmids.

A large number of plasmids have been identified by chance during the analysis of host bacteria based on specific phenotypes (8). However, this method of identification is very laborious and cannot provide proper insight into the so-called 'plasmidome', which refers to the entire plasmid DNA content of a particular environmental sample independent of cultivation (3). Although culture-dependent plasmidome studies (9–12) have significantly contributed to our understanding of mobile genetic elements from a given bacterium or bacterial group, plasmids from the non-cultivable organisms are out of reach using these methods. Several approaches addressing this issue have been developed, including exogenous plasmid isolation (13,14) and Transposon-aided Capture (TRACA, (6,15–16)), but these methods have significant limitations. For low-throughput exogenous plasmid isolation, the plasmid which is to be captured needs to be conjugative (or at least mobilizable) and stably replicated in the recipient cell (3). This significantly reduces the number of possible hits and misses of plasmids that are not mobile. In contrast, TRACA is more high-throughput, but can only capture small plasmids of 2–10 kb in size (3,17). Although direct isolation of plasmids from the environment is possible, it is generally restricted to small ones (17). A technique using indirect plasmid isolation from an environmental sample followed by the exonu-

*To whom correspondence should be addressed. Tel: +48 22 592 20 30; Fax: +48 22 658 4176; Email: p.krawczyk@ibb.waw.pl

clease treatment and Phi29 amplification was previously developed (18–20) but it is more useful for samples with high bacterial biomass content and a low amount of contaminants interfering with enzymatic procedures.

Bioinformatic methods for the identification of plasmid sequences in metagenomic datasets have also been developed but, analogous to molecular methods, they are generally aimed at the identification of circular elements (21). Fast development of sequencing technologies and reduction of sequencing costs has led to an increase in the number of metagenomic projects, resulting in the constant growth of sequence databases. As a given metagenome is mostly a mixture of chromosomes and plasmids in which the relation of both vastly remains in favor for the chromosomal (3), many plasmid sequences have been included in sequenced metagenomes but remained unidentified. Some attempts have been made to assemble plasmids from metagenomic data (6,22) but they rely on laborious and computationally intensive approaches. Recently, the SPAdes assembler version capable of assembling plasmids was developed (23), as well as the Recycler, aimed at recovering circular contigs from de Bruijn assembly graphs (24). Another simple approach is implemented in the PlasmidFinder (25), which is a web-based program aimed at identifying plasmid replicons in bacterial genome assemblies. This program conducts a similarity search against well-defined replicon sequences to identify possible plasmids. Although it can predict plasmid-originated sequences with high precision, its main limitation lies in the size of the database, which is composed mostly of *Enterobacteriaceae* replicons. This aspect significantly limits its usage for metagenomic studies. A database similarity search approach is extended in the Plasmid Constellation Network (PLACNET), which uses BLAST to compare sequences against reference databases and then a network analysis to reconstruct plasmids (26). However, this program relies on the manual curation of obtained sequence clusters, thus preventing its use in any automatic annotation pipeline. Moreover, the results obtained from PLACNET are not fully reproducible and depend on the expertise of the researcher.

Another interesting approach for identification of plasmids from shotgun metagenomic data is the use of machine learning techniques which should allow for the automatic detection of plasmid sequences in any given metagenome. Genome signature-based methods have revealed that plasmid-host similarity correlates with genomic %G+C content (27,28). The barcodes of plasmid genomes also tend to have similar characteristics, possibly due to similar selection pressure caused by their frequent transfer among cell cultures (29). A method for the identification of plasmid sequences based on genomic signatures was implemented in the cBar software (30).

Recently, the performance of available tools aimed at plasmid reconstruction from sequencing data was compared (31) and showed that although such an automatic procedure is possible, there are still limitations, especially for large (>50 kb) plasmids. However, that review focused on single bacterial isolates rather than complex sequencing data from metagenomic projects. Consequently, there is a lack of systematic comparison of algorithms oriented towards plasmid identification in metagenomic datasets. In this study, we present PlasFlow, a novel approach for the prediction of bacterial plasmid sequences in metagenomic contigs and compare it to other available tools. Using genome signatures of sequences from 9,565 bacterial chromosomes and plasmids, we trained a deep neural network model to separate chromosomal and plasmid sequences from different phyla. Our approach achieved as much as 96% classification accuracy in plasmid prediction on test data, which is significantly better than any other previously developed tool. Tests performed on real metagenomic datasets revealed its versatility for plasmidome analyses using assembled metagenomic data.

MATERIALS AND METHODS

Datasets

To train the model, we used 9,565 fasta sequences, including both chromosomes (1961) and plasmids (7604) of organisms from the kingdom Bacteria, which we downloaded from the NCBI Refseq Genomes FTP (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq) based on the following criteria: (i) genome was marked as the 'representative genome'; (ii) genome was at the assembly level 'Complete genome'; (iii) for a given species, only the most up-to-date sequence was downloaded. Additionally, plasmid sequences were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid/. Taxonomic information was obtained using the rentrez (Winter, 2016, available at https://CRAN.R-project.org/package=rentrez) package in R based on taxon_id provided in the assembly_summary.txt file from the NCBI FTP, and were further manually curated to filter out eukaryotic, archaeal and viral sequences. Training sequences were then deduplicated to remove plasmids occurring both in the whole genome sequences and in the Refseq plasmids database. All full-length sequences used for training are listed in Supplemental Table S1. In the next step, all sequences were randomly split to smaller fragments with lengths of 5, 10 or 50 kb, which covered up to 5 or 75% of the sequence length of individual chromosomes and plasmids, respectively. The remaining sequences, which were not included in the 10 kb training dataset, were filtered to have a minimal length of 1 kb and used for further performance analyses. This validation dataset contained 61 221 sequence fragments with lengths ranging from 1 to 1580 kb. All sequence fragments were labelled using origin (plasmid/chromosome) and taxonomic (phylum) information (see details below).

To test our approach, we used reference plasmidome data, which included: (i) the bovine rumen plasmidome assembly (Brown Kav dataset) (18), which was downloaded from the MG-RAST (32) (accession 4460391.3); (ii) the plasmid metagenome of an activated sludge system assembly (33), which was downloaded from the MG-RAST (accession 4474000.3); (iii) the plasmid metagenome of an activated sludge (Zhang dataset) (6), which was downloaded as raw reads from the NCBI Short Read Archive (SRA) (accession SRP007256.1) and assembled using SPAdes 3.9.1 (34); (iv) plasmids from the wastewater treatment plant (Szczepanowski dataset) (12) which was downloaded from ftp://ftp.cebitec.uni-bielefeld.de/pub/supplements/SzczepanowskiEtAl_Insight_JournalBiotech_2008.zip as

raw reads and assembled using SPAdes, with assembler-only option due to lack of quality data.

Applicability of presented software to newly sequenced genomic data was assessed using recently published *Aeromonas* sp. 023A (35), assembled using SPAdes, and draft genomic sequences available in the Refseq database, matching following criteria: (i) the genome was marked as the 'representative genome'; (ii) the genome was at the assembly level 'Scaffold' or 'Contig' and (iii) for a given species, only the most up-to-date sequence was downloaded. An additional test was performed using a set of 42 bacterial genomes that contained a various number of plasmids, which has been described in Arredondo-Alonso *et al.* (31). A list of genomes and their accessions is provided in Supplemental Table S2. Raw sequencing data for each genome were downloaded from SRA and assembled using SPAdes 3.9.1. Reference assembly sequences were downloaded from GenBank.

Evaluation of metagenomic data was performed using sequences from microbial mats inhabiting mine waters from an abandoned uranium mine in Kowary (KOW) and a gold mine in Zloty Stok (ZS), Poland. Raw reads, available at the MG-RAST (accession numbers in Supplemental Table S3), were assembled using SPAdes 3.9.1 (34).

Sequence processing and neural network training

A flowchart illustrating major steps concerning preparation of the training dataset, neural network training, and sequence classification is presented in Figure 1.

Sequence fragments from the training dataset were arranged in classes containing information about the sequence origin (plasmid or chromosome) and taxonomic classification (at the level of phyla). If the number of sequence fragments in a given class was <100 (due to a low number of representatives in the database), all sequences from such a class were grouped as chromosome.other or plasmid.other, to avoid weakly represented classes in the input of the neural network. For the 10 kb dataset chromosomal sequences from phyla *Aquificae, Caldiserica, Chrysiogenetes, Deferribacteres, Dictyoglomi, Fibrobacteres, Elusimicrobia, Ignavibacteriae,* and *Thermodesulfobacteria* and plasmid sequences from phyla *Acidobacteria, Aquificae, Chlorobi, Chloroflexi, Deferribacteres, Elusimicrobia, Nitrospirae, Planctomycetes, Synergistetes, Tenericutes, Thermotogae* and *Verrucomicrobia* were excluded.

Genomic signatures were represented as the vector of frequencies of all oligonucleotides (kmers) of desired length (k) occurring in an analyzed sequence (s), defined as a string over the alphabet {a, t, c, g}. The total number of possible kmers of length k is given by 4^k, e.g., for $k = 3$, the genomic signature is the vector of 64 elements, and, for $k = 4$, the genomic signature is the vector of 256 elements.

Kmers of 3–7 nt were counted using the function OligonucleotideFrequency from the BioStrings 2.46 (Pages *et al.*, available at http://bioconductor.org/packages/release/bioc/html/Biostrings.html) package in R, then transformed using the Term Frequency-Inverse Document Frequency (TF-IDF) method from the SciKit-learn 0.18 python package (36), resulting in normalized kmer frequencies. TF-IDF transformation was used to minimize the impact of frequently occurring kmers on subsequent calculations. The TF-IDF value for kmer t is calculated as tf-idf(s, t) = tf(s,t) * idf(t), where tf(s,t) is the kmer t frequency in the sequence s, and the idf is computed as idf(t) = log [n/df(s, t)] + 1, where n is the total number of analyzed sequences and df(s, t) is the number of sequences s that contain the kmer t. TF-IDF vectors for each sequence were normalized using Euclidean (L2) normalization (as implemented in SciKit-learn).

To properly monitor the progress of neural network learning and accurately calculate metrics measuring the performance of trained neural networks, we used part of the training dataset in the cross-validation procedure. We split obtained datasets containing predictors (vectors of TF-IDF transformed and normalized kmer frequencies) and target variables (labels of classes) into two datasets containing 75% and 25% of all data, respectively, using the train_test_split function from the Scikit-learn with the option 'stratified' to maintain the structure of classes. The first dataset was then used for neural network training using the TensorFlow's 0.10.0rc0 (Abadi *et al.*, available at https://www.tensorflow.org) high-level machine learning API (tf.contrib.learn). The second dataset was used for the in-training evaluation of learning procedure using TensorBoard (which is a part of TensorFlow).

Training, using the ReLu activation and the AdaGrad optimizer, was conducted in 50,000 steps with the following measurements (based on the aforementioned testing dataset): accuracy (Ac), precision (Pr), and recall (Re), which were saved every 100 steps to evaluate the classification accuracy of a model. For each kmer length, we trained neural networks composed of one hidden layer with 20 or 30 neurons or two hidden layers with 10 or 20 neurons each. Final models (after 50 000 learning steps) were evaluated using the SciKit-learn python package, calculating accuracy (Ac), f1 score, precision (Pr) and recall (Re).

Accuracy is the fraction of correct predictions, calculated as (tp+tn)/n, where tp is the number of true positives, tn is the number of true negatives, and n is the number of all samples. Precision = tp/(tp + fp), where tp is the number of true positives and fp is the number of false positives. Recall = tp/(tp + fn), where tp is the number of true positives and fn is the number of false negatives. The f1 score is calculated as: f1 = 2 * (precision * recall)/(precision + recall).

Metagenomic assembly

Metagenomic shotgun reads were adapter trimmed using cutadapt 1.9.2 (37), then filtered by quality (minimal quality 30, minimal length 50) using sickle 1.33 (Joshi *et al.*, available at https://github.com/najoshi/sickle) and assembled using SPAdes 3.9.1 (34). Reads used for assembly were deposited in the MG-RAST database (32) at the accession numbers listed in Supplemental Table S3.

Classification using PlasFlow

As short sequences cannot be efficiently used in genome signature calculations (not including reference plasmidomes), we filtered out sequences that were shorter than 1 kb. Filtered sequences were then used in the kmer counts calculation (kmers length 5–7 nt) using the function OligonucleotideFrequency from the BioStrings package in R, then

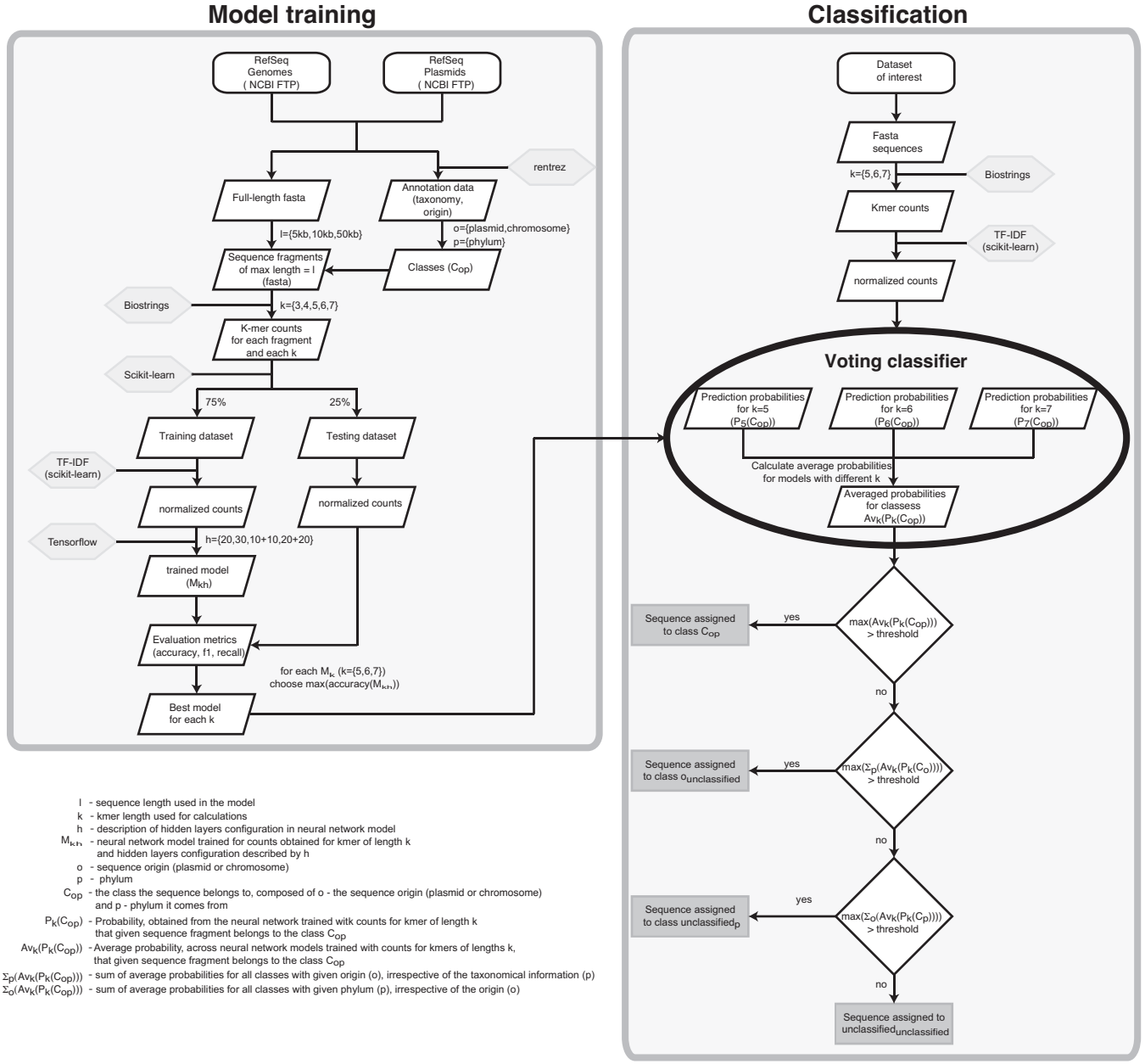

Figure 1. Flowchart describing the training and classification procedures implemented in the PlasFlow.

transformed and normalized using the TF-IDF method as described above. Normalized kmer frequencies were used for classification using the obtained TensorFlow classifiers (separate for each kmer length used). A voting classifier was built using the best models for kmers of length 5–7 nt, in which probabilities of assignment to each of an allowed class were averaged over the three classifications to obtain the final classification score (Figure 1). The average score ($Av_k(P_k(C_{op}))$ was calculated as $\Sigma(P_k(C_{op}))/3$, where C_{op} is a class containing information about origin (o) and phylum (p) and $P_k(C_{op})$ is a probability of assignment to class C_{op} returned by the best classifier trained on kmers of length *k* (for *k* in the range 5–7 nt). The sequence was assigned to a class with $\max(Av_k(P_k(C_{op})))$ but if the resulting maximum average score was lower than the specified threshold (default value in this paper = 0.7), then a sequence was assigned as 'unclassified.unclassified,' meaning that neither plasmid nor phylogenetic classification could be conducted on a given sequence. However, to allow for broad host-range classification, if a sum of average probabilities for all plasmid classes, irrespective of the taxonomical information, was higher than the threshold, then the sequence was assigned as 'plasmid.unclassified', meaning that the signature is somehow similar to that of plasmids but we cannot say from which taxonomic group it came. Similarly, if a sum of probabilities for all chromosome classes, irrespective of the taxonomical information, was higher than the threshold, then the sequence was as-

signed as 'chromosome.unclassified'. Finally, if a sum of probabilities for plasmid and chromosome sequences of a given phylum was higher than the threshold, then the sequence was assigned as 'unclassified.phylum', where 'phylum' is the name of the phylum with the summary probability higher than the threshold. This allowed for taxonomical prediction in the case where plasmid predictions were ambiguous.

Prediction of plasmid sequences using cBar

cBar 1.2 (30) was downloaded from http://csbl.bmb.uga.edu/~ffzhou/cBar/. Sequences were filtered by length analogously to the PlasFlow classification (described above) and used as an input to cBar. Obtained classification file was then used to extract sequences assigned to plasmid or chromosome class using prepared Perl script.

Prediction of plasmid sequences using PlasmidFinder

To predict plasmid replicons using PlasmidFinder (25) we uploaded assembled sequences (filtered by length—as described above) to the PlasmidFinder webserver (https://cge.cbs.dtu.dk/services/PlasmidFinder/) and ran the computation selecting all available databases (*Enterobacteriaceae* and *Enterecoccus*, *Streptococcus*, *Staphylococcus*). The %ID threshold was set at 80% and 'Assembled Genome/Contigs' was chosen as the type of read. Results were downloaded as raw text files.

Prediction of plasmid sequences using Recycler

Recycler 0.61 (24) was downloaded from https://github.com/Shamir-Lab/Recycler. The SPAdes assembly graph fasta file was created using make_fasta_from_fastq.py script available in Recycler. The BAM file required as input by Recycler was created by the alignment of the reads used for the assembly against the created assembly graph fasta file using Bwa 0.7.15 (38) and samtools 1.4–22 (39). Cycles in the assembly graph identified by Recycler were considered to be possible plasmids.

Evaluation metrics for genome assemblies

To compare different software performance on single genome assemblies, we used the same dataset and metrics as described previously (31). Scaffolds obtained from the SPAdes assemblies were filtered by length (as described above) and used for plasmid prediction by PlasFlow or cBar. Due to a lack of the length filtering step in the Recycler or PlasmidSpades, those were excluded from the current analysis since it was already performed in the aforementioned publication. Plasmid bins from both PlasFlow and cBar were compared to the reference sequences using Quast 4.3 (40). Recall was defined as the fraction (in terms of length) of reference plasmid(s) covered by the prediction, whereas precision was calculated as tp/(tp+fp), where tp represents the overall contig length for true positives (true plasmid predictions) and fp represents the overall contig length for false positives (plasmid predictions which are coming from chromosome sequences).

Sequence annotation and statistical analysis

Sequence bins obtained from the metagenomic assemblies using different classification approaches were annotated by the mean of Subsystems using MG-RAST pipeline (32). Functional profiles were downloaded from MG-RAST at the function level. For the Principal Component Analysis (PCA) counts for level1 subsystems were summarized, normalized to the sample size, and analyzed using prcomp function from R to reduce dimensionality and visualize differences between individual samples and groups. Chromosomal, plasmid and unclassified sequence annotations for each metagenome as well as chromosomal and plasmid sequence annotations from different metagenomes were compared to each other using standard options of the DESeq2 package (41) in R.

RESULTS

Rationale and design

Nucleotide composition, usually measured as the %G+C content, is a simple and the most commonly used metric to describe sequences. There are clear differences in %G+C base pairs between species and very low intra-species variance, making it a useful parameter for taxonomic classification or genome binning (42). However, whole genomes are usually compared without distinguishing chromosomal sequences from plasmids. There are only a few reports comparing the %G+C content of plasmids and their hosts' chromosomes (43–45). A fast increase in the number of bacterial genomes sequenced makes such comparisons more reliable. Therefore, we evaluated sequence statistics of 1961 and 7604 full-length chromosome or plasmids sequences, respectively, obtained from the NCBI RefSeq Genomes database (Supplemental Table S1).

Analyzed chromosome sequences had a mean %G+C of 51.05% and a mean length of 3587 kb (range 83–14 782 kb), whereas plasmids had a mean %G+C of 44.31% and a mean length of 85 kb (range 0.4–2658 kb). These results are in agreement with the findings of Shintani *et al.* (8) who analyzed 4602 plasmid sequences with a mean size of 80 kb and average %G+C content of 44.1%. Although mean plasmid length is lower, as the %G+C content, there is only a weak correlation between %G+C content and sequence length (Pearson's $r = 0.39$, Supplementary Figure S1) and there are clear differences in the mean %G+C content between phyla.

The lowest %G+C in the group of plasmids was observed for Fusobacteria, 25.95% ($N = 19$, mean length 71 kb), and Tenericutes, 26.12% ($N = 65$, mean length 6 kb). Fusobacteria and Tenericutes had also the lowest chromosome %G+C content (29.55%, $N = 8$, and 28.50%, $N = 56$, respectively). The highest mean %G+C content of plasmid and chromosomal DNA was observed for phylum *Deinococcus-Thermus* (64.92%, $N = 45$ and 67.18%, $N = 23$, respectively). %G+C content of sequences belonging to the same phylum were similar and there were apparent differences in the %G+C content between phyla (Supplementary Figure S1). The mean %G+C content of plasmids and chromosomes from the same phylum differed significantly in most cases and the mean %G+C content of plasmids was usually

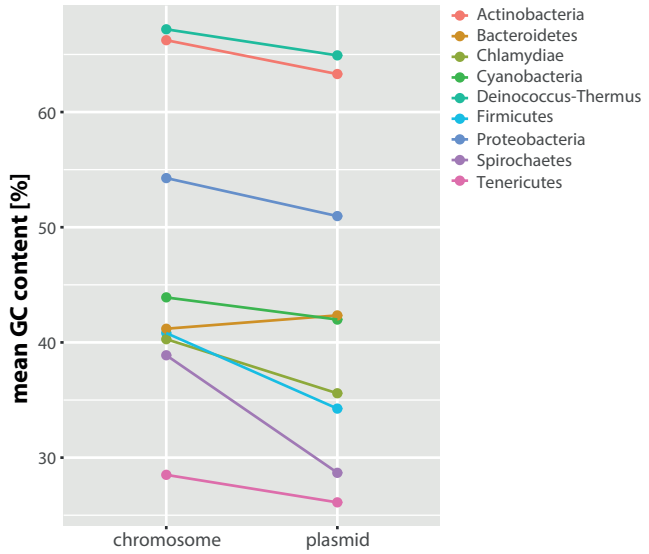

Figure 2. Comparison of mean %G+C content of chromosomes and plasmids grouped by phyla. Only phyla with at least 10 sequences in a group of plasmids or chromosomes shown.

lower than that of chromosomes (Figure 2); this is consistent with findings of Nishida *et al.* (43).

Since the %G+C content analysis of chromosomes and plasmids revealed significant differences between plasmid- and chromosome-derived sequences and a clear separation between taxons (phyla), we investigated if more detailed sequence characteristics could be used to predict if a given sequence comes from plasmid or chromosome. We postulated that a machine learning approach based on carefully selected genome signatures represented as kmer frequencies may provide reasonable results, as such an approach has successfully been used for genome comparisons in previous studies (46–49). Machine learning methods are receiving much attention since a rapid increase in biological data dimensions is challenging conventional analysis methods (50). Deep artificial neural networks are becoming state-of-the-art in predictive analyses, as they allow for efficient exploitation of complicated data, discovering high-level features and providing an additional understanding about the structure of biological data (50). For the purpose of analyses described herein, we used the TensorFlow framework, recently released by Google (Abadi *et al.*, available at https://www.tensorflow.org), allowing for easy implementation of deep neural networks for various complicated datasets, including biological data (50–53). Major steps of data analysis are presented as a flowchart (Figure 1). Genomic signatures, which we used for neural network training, were calculated for the bacterial chromosome and plasmid sequences downloaded from the NCBI RefSeq (Supplemental Table S1) using kmers with lengths in a range 3–7 nt, as these are known to be the most informative (54). To reliably train the neural network, the number of samples used for training should be significantly larger than the number of predictors (individual kmer frequencies in this case). As for heptamers, we achieved 16 384 predictors (4^7), which was much more than the number of full-length sequences (9565); therefore, we decided to randomly split all sequences into smaller fragments. As a default, we decided to choose a length of 10 kb (or less for short plasmid sequences present in the dataset), but fragments of length 5 and 50 kb were also tested. Since the average length of chromosomes is higher than that of plasmids, the number of chromosome sequences after the splitting procedure would be significantly higher than that of plasmid sequences in the training dataset, which could bias neural network training. Therefore, we decided to use only randomly selected fragments, which covered up to 5% or 75% of sequence length of individual chromosomes and plasmids, respectively. In this way, we obtained a proper balance of classes and increased the number of samples (89 509 for 10 kb dataset) to a level much higher than the maximum number of possible kmers we can observe (for $k = 7$, $4^7 = 16\ 384$). In addition, we found such splitting reasonable because metagenomic assemblies are generally fragmented with a mean contig length and N50 around 1–2 kb (55). Therefore, the training dataset comprised of shorter sequences was similar to a typical metagenomic dataset.

As high variance in %G+C between phyla was revealed (Figure 2), we grouped training sequences in the taxonomic bins labelled with the origin (plasmid or chromosome) and phylum from which they are derived, so that, in the final classifier, we are able to infer not only origin (plasmid or chromosome) of sequence but also minimal taxonomy information. The TensorFlow framework was applied to obtained data using different configurations of hidden layer neurons, including 1- and 2-layer design to train the network. For the all fragmentation schema tested, the best prediction accuracies were observed for kmers with lengths 5 to 7 nt (Figure 3A; Supplemental Table S4). Multiple models displayed high prediction accuracy (in a range of 80–90%), indicating that our approach was indeed highly reliable. Since the performance of models trained using kmers of $k = \{5–7\}$ was similar, we decided to build a voting classifier that uses the average of prediction probabilities obtained from the best models trained using aforementioned lengths of kmers (Figure 3A, marked in red, for 10 kb sequence fragments) to draw the final prediction. Such ensemble methods are usually used to improve the performance of individual classifiers, as they allow elimination of classification errors made by single classifiers (56). The typical output of a neural network-based classifier is a vector containing probabilities of an assignment to a given class, which is then translated to class labels based on the highest probability. In case of ensemble method like in our approach, a result is the vector containing probabilities averaged over all applied classifiers. An element may be classified in a given class even if there is a low probability of such classification that may lead to erroneous predictions. Therefore, to avoid probable misclassifications, we introduced a threshold, the minimal average probability required to treat the sequence as classified to a given class. We assumed that classification would be correct if at least two out of three classifiers assigned a sequence to a given class with a high probability and thus used a default probability threshold of 0.7. Sequences that did not fall above the specified threshold were treated as unclassified because predictions cannot be drawn with high certainty. Because our approach allowed for prediction of both origin of sequence (plasmid or chro-

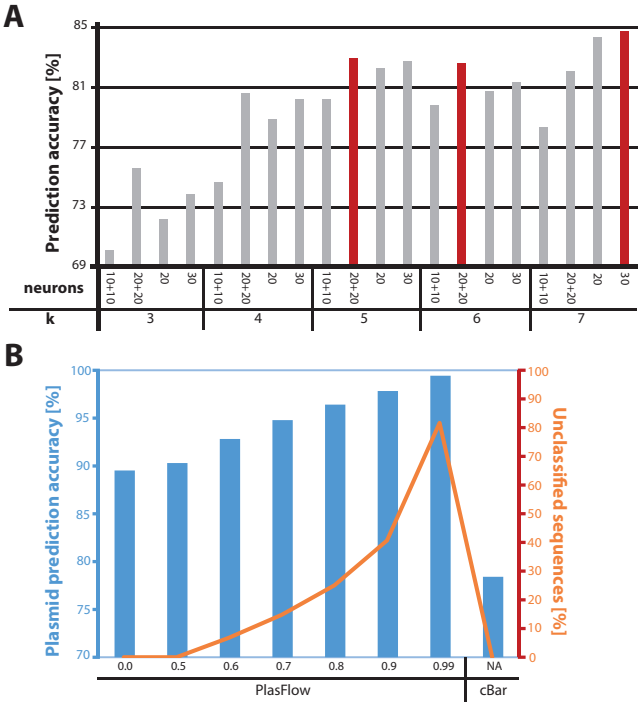

Figure 3. (**A**). Prediction accuracies obtained from TensorFlow-based classifiers trained on 10 kb sequence fragments using different input data and hidden layers configuration. As the input to the network, kmer frequencies were calculated using kmer lengths in range $k = \{3,4,5,6,7\}$. Hidden layer configurations tested included one-layer design with 20 or 30 neurons or a two-layer design with 10 or 20 neurons in each layer. Classifiers with the best accuracies, used in the PlasFlow classifier, are marked in red. (**B**) Performance of PlasFlow classification on fragmented RefSeq genomes and plasmids using different probability thresholds. Plasmid prediction accuracy is shown in bar chart; number of unclassified sequences using a given threshold is shown in line chart. Respective data for cBar shown for comparison.

mosome) and taxonomic classification, we decided to modify final classification to allow for assignment of a sequence to the group of plasmids or chromosomes, even if an accurate taxonomic assignment was impossible; this way, the origin can be predicted for sequences harboring properties from different taxonomic groups, as in the case of the broad-host-range plasmids.

In order to verify the performance of the ensemble classifiers trained on sequence fragments of length 5, 10 or 50 kb, they were tested on the validation dataset containing sequences with lengths ranging from 1 to 1570 kb and composed of fragments of sequences listed in Supplemental Table S1. Without probability filtering, classifier trained on 10 kb fragments obtained 89.52% accuracy for plasmid classification; introducing filtering at 0.7 improved the accuracy to 94.79%, with a low number of sequences unclassified (Figure 3B; Supplemental Table S5).A higher probability threshold (0.8) increased accuracy to 96.41%, with 25.03% of sequences unclassified. The same tests performed with ensemble classifiers trained on 5 or 50 kb fragments indicated that although prediction accuracies are generally on a similar level, they perform significantly worse when we consider a fraction of unclassified sequences and the false positive rate (Supplemental Table S5). It is likely that the kmer coverage for the 5 kb fragments is too low (especially for longer kmers) and therefore some predictions fail. In the case of 50 kb fragments, for which we obtained the best scores during the initial evaluation of models (Supplemental Table S4), a high number of false positive results may indicate a possible overfitting of the neural network. According to the results of this benchmark, we decided to continue using models trained on 10 kb sequence fragments and included them in the newly developed software, PlasFlow.

When comparing the PlasFlow to cBar, a previously described software for the prediction of plasmid sequences in metagenomic datasets (30), PlasFlow performed significantly better, reaching an order of magnitude higher accuracy of plasmid prediction (Figure 3B). Due to probability filtering, PlasFlow outputs a significantly lower number of false predictions than cBar (Supplemental Table S5), what additionally proves its reliability in plasmid prediction. In the analyzed dataset, cBar incorrectly predicted 6.46% of sequences to be plasmids (false positives) and as many as 15.14% to be chromosomes (false negatives), whereas for PlasFlow with 0.7 threshold, those numbers were much smaller (2.21% false positive and 3.00% false negative predictions) (Supplemental Table S5). It should be noted that 79% of predictions were consistent between cBar and PlasFlow (run without any probability filtering).

In summary, we developed a training and classification procedure for prediction of plasmid sequences with unprecedented accuracy and a low false positive error rate. The algorithm is summarized in the flowchart shown in Figure 1. Briefly, reference chromosomal and plasmid sequences were fragmented and used for calculation of genome signatures, which were then provided as an input for neural network training. Out of several models obtained we selected those that performed best for kmers of length 5, 6 or 7 nt and used them to build the voting classifier, which executes the actual prediction of plasmids. Additional thresholding was then introduced at this step, which allows for the exclusion of erroneous predictions and significantly improves the overall accuracy of classification. We combined all required calculations into a simple Python script, which accepts fasta file format as an input and returns a table with classification results as well as fasta files with sequences assigned to the group of plasmids, chromosomes, and unclassified sequences, respectively. The default probability threshold was set at 0.7, but users can specify their own filtering. The software, PlasFlow, is available at GitHub (https://github.com/smaegol/PlasFlow).

Performance on public plasmidome datasets

Next, we asked if PlasFlow could successfully classify sequences in metagenomic datasets with a well-defined composition for plasmids. For this purpose, we used publicly available plasmidome datasets from experiments in which metagenomic DNA was enriched in plasmids using exonuclease treatment (in case of Brown Kav, Zhang datasets) (6,18) or cesium chloride gradient centrifugation (in case of Sentchilo and Szczepanowski datasets) (12,33). Techniques applied during the preparation of those datasets were aimed at removal of linear DNA fragments, leaving only circular elements, a gross majority of which came

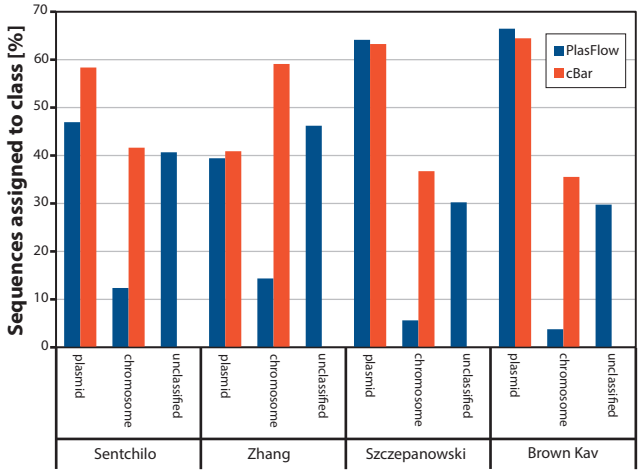

Figure 4. PlasFlow evaluation on public plasmidome data. For each assembled plasmidome dataset, classification was performed using PlasFlow (with a 0.7 probability threshold) and cBar. For each dataset, the percentage of sequences classified as chromosomal, plasmid or unclassified is shown.

from plasmids. Most of the analyzed sequences were predicted to come from plasmids (Figure 4), what can be clearly seen when we analyze the ratio of number of identified plasmids to the number of identified chromosome sequences (Supplemental Table S6). For the Brown Kav and Szczepanowski datasets, we obtained plasmid to chromosomes ratios 15.73 and 13.55, respectively, while for the Sentchilo and Zhang, we obtained plasmid to chromosomes ratios 4.72 and 2.83, respectively. A significant number of sequences remained unclassified due to low probabilities (threshold = 0.7) obtained from the classifier, which could have been a result of low lengths of sequences in tested datasets. Compared to cBar, which obtained plasmid to chromosome ratios of 1.81, 1.40, 1.72 and 0.69 for the Brown Kav, Sentchilo, Szczepanowski and Zhang datasets, respectively, PlasFlow allowed for significantly better separation of plasmid and chromosomal sequences. Furthermore, 44% of predictions were shared between PlasFlow and cBar, 85% of which came from plasmids, indicating that PlasFlow was able to exclude most of the unreliable chromosomal predictions of cBar. Notably, PlasmidFinder, which uses a completely different BLAST-based approach for plasmid identification, only identified plasmid sequences in the two datasets (5 contigs for Sentchilo and 15 contigs for Szczepanowski datasets), which underscores its limitations for metagenomic assemblies. Other plasmid prediction software, like Recycler or PlasmidSpades, could not be used for this benchmark due to a lack of the required pair-end sequencing data.

It should be noted that, for the two analyzed datasets (Brown Kav and Zhang), authors checked the quality of plasmidome enrichment using PCR amplification of 16S rRNA gene, which is generally absent from plasmid sequences (57–59). For the Brown Kav dataset, authors were not able to find any 16S rRNA amplification products (18), whereas for the Zhang dataset, significant chromosomal contamination was found (6). These results can explain lower plasmid enrichment in the Zhang dataset in both PlasFlow and cBar predictions. These results also show the limitations of molecular biology methods used for plasmidome analyses, which cannot effectively remove all chromosomal contaminations, which may be one of the reasons why we were not able to classify all of the sequences in the analyzed plasmidome datasets as plasmids, even if they were screened for contamination using 16S rRNA amplification.

Applicability for genome assembly

Although we aim the usage of PlasFlow for the analysis of metagenomic datasets, it can be also easily applied to single genome assemblies, allowing for identification of possible plasmids present in analyzed genomes. PlasFlow can be especially beneficial at the early steps of genome assembly, when usually a high number of contigs is obtained, which require further processing steps, including scaffolding and gap closing. Identification of contigs, which should be scaffolded separately from the chromosome, may reduce the time needed for assembly finishing and provide information about the genome structure of the sequenced organism.

We assessed the performance of PlasFlow using the recently published *Aeromonas* sp. O23A genome, which contains four plasmids with size ranging from 4 to 60 kb (35). Preliminary genome assembly was performed using SPAdes, resulting in 240 scaffolds, which were further filtered to exclude all sequences shorter than 1 kb. Out of the remaining 36 sequences, PlasFlow analysis identified eight potential plasmid sequences (Supplemental Table S7) that we compared to the manually curated *Aeromonas* sp. 023A plasmid sequences using nucleotide BLAST. We found significant matches to the all known *Aeromonas* sp. 023A plasmids in 5 out of 8 sequences. At the same time cBar was also able to correctly predict all plasmid sequences.

To further show the applicability of PlasFlow for plasmid sequence separation in genome assembly projects, we applied our solution to the set of 42 bacterial genomes described in the recent comparison of methods aimed at reconstruction of plasmids from whole-genome sequencing data (31). To measure PlasFlow performance we used the same metrics as the authors of the aforementioned publication and found that PlasFlow outperforms the other described methods (Supplemental Table S2), since it recovered 85.98% (recall) of plasmid sequences present in analyzed samples compared to 76.82% in the case of cBar and 12.0% for Recycler (31). Moreover, PlasFlow also shows higher precision (72.17% versus 60.51% for cBar and 30.00% for Recycler), indicating that it is able to more efficiently exclude false positive hits.

Genome sequence databases are constantly growing due to a high number of projects involved in the sequencing and assembly of bacterial genomes. However, a high number of assembled genomes never reach the 'Complete genome' stage and remain in the form of contigs and scaffolds, without any information about the genome structure. Therefore, we evaluated if PlasFlow could identify plasmid sequences in unfinished genome sequences.

For this purpose, we used sequences from the NCBI Refseq with assembly status 'Contig' or 'Scaffold' and marked as 'representative genome'. For analysis, we used 101 454 sequences of length >10 kb. PlasFlow classification of se-

lected sequences yielded 6753 (6.66%) plasmid, 78 292 (77,17%) chromosome and 16 403 (16.17%) unclassified fragments (Supplemental Table S8). In 89 502 cases (88.22% of all analyzed sequences) the phylum was properly assigned. Similar approaches can be used for a more detailed taxonomic classification, however, this was out of scope of our research.

Since it was not possible to unambiguously confirm PlasFlow predictions due to a lack of proper experimental data, we assessed the accuracy of predictions using available annotation data. We verified analyzed sequences with genes found on plasmids, coding conjugal transfer proteins, plasmid partitioning proteins, plasmid replication initiation proteins, or plasmids stabilization proteins. Using annotations available at Genbank, we found that 920 (13.61%) and 2722 (3.48%) of sequences classified as plasmids or chromosomes, respectively, have such annotation (Supplemental Table S8).

Although there is a marked increase in the plasmid-related annotations in sequences classified as plasmids, it is apparent that the majority of sequences do not possess such annotations. However, it should be noted that there are a large number of plasmids that do not code any proteins responsible for conjugative transfer and are only transferred vertically. Since the presence of genes responsible for the mobility of plasmids is not necessary, it is very difficult to predict if a sequence is plasmid based only on the functional annotation data.

Performance on environmental metagenomic datasets

The most desirable feature of PlasFlow would be applicability to metagenomic data, which should allow for comparative plasmidome analyses without the need for laborious laboratory techniques; such an approach should permit detailed description of plasmids contribution to microbial communities from diverse environments.

To evaluate PlasFlow applicability in comparative plasmidome analyses, we analyzed metagenomes from microbial mats inhabiting mine waters from the abandoned uranium mine in Kowary (KOW) and the gold mine in Zloty Stok (ZS), Poland, which were observed in the bottom sediments of both slightly acidic and neutral mine waters with elevated concentrations of heavy metals. Both mines were previously described as rich in functions related to heavy-metal homeostasis (60) but they were analyzed only as raw reads without the assembly step. Thus, for the purpose of plasmidome analyses, we assembled 3 independent samples from each environment using SPAdes 3.9.1 (statistics of assembly are available in Supplemental Table S3). Obtained scaffolds were filtered to the length of >1 kb and provided as an input to PlasFlow, cBar and PlasmidFinder for plasmid prediction. Although PlasFlow (similar to cBar) is based on completely different assumptions, we decided to compare it to the Recycler, which identifies cycles in the assembly graph that represent possible circular plasmids. We did not consider the PlasmidSpades, which focuses mainly on identifying plasmids from sequencing data of genomes obtained from cultivated bacteria (23). Sequences classified into chromosome and plasmid bins (or just possible plasmids in the case of Recycler) were then annotated using the MG-RAST server (32) to check if we were able to see enrichment in plasmid-related functions in sequences classified as plasmids. We used subsystem classification profiles downloaded from MG-RAST for the purpose of statistical analysis and identification of functions enriched in obtained sequence bins.

For all analyzed samples, PlasFlow identified 25.02–27.05% of sequences to be of plasmid origin that constituted 14.94–20.66% of cumulative scaffolds length (Supplemental Table S9). High numbers of sequences remained unclassified due to low probabilities obtained from classifier. Principal Component Analysis (PCA) of analyzed samples revealed that chromosomal sequences from both metagenomes were similar to each other, whereas plasmid and unclassified differed from them as well as from each other (Figure 5). First principal component (PC1), which best resolved the variability between chromosomal, plasmid, and unclassified bins, explained 44.5% of variance between samples. Importantly, the Phages, Prophages, Transposable Elements, and Plasmids subsystem contributed mostly to that component. This is in concordance with the observation that the most pronounced enrichment in the group of plasmids was visible for this subsystem (2.99% and 4.10% for plasmids and 1.16% and 1.14% for chromosomes in KOW and ZS samples, respectively). More detailed analyses at deeper subsystems levels (Supplemental Figure S3, Table S10) revealed that Phage integration and excision subsystem contributed mostly to this enrichment (1.36% and 1.02% for plasmids and 0.28% and 0.22% for chromosomes in KOW and ZS samples, respectively), which could be due to two reasons. First, it is possible that phage and plasmid sequences are similar (by the mean of nucleotide composition) and PlasFlow cannot differentiate between them. Second, the annotation may be imprecise, whereby it could describe Integrative Conjugative Elements (ICEs) as phages, although they are commonly found on plasmids. This is in agreement with the review reporting that tyrosine integrases of ICEs are frequently annotated as 'Phage Integrases' (61), and such were also frequently found in the annotation data of our datasets (Supplemental Figure S3, Table S10). Although we hypothesize that the latter explanation is correct, definitive conclusions cannot be drawn without additional experimental effort.

Another subsystem for which we see enrichment in the plasmid dataset is the Toxin-antitoxin systems (other than RelBE and MazEF) (constituting 0.50% and 0.91% in plasmids and 0.16% and 0.22% in chromosome-assigned sequences in KOW and ZS samples, respectively). This subsystem groups mostly the small proteins from the type II toxin-antitoxin systems involved in the stabilization of plasmids in the host cells (62). Importantly, we found that the protein and nucleoprotein secretion system, Type IV subsystem, that groups proteins involved in the conjugative transfer was also overrepresented in the plasmids dataset when compared to chromosomal one (0.50% and 1.19% sequences assigned to plasmids and 0.11% and 0.12% assigned to chromosomes in KOW and ZS samples, respectively, contained annotations from this subsystem), proving that most conjugative plasmid elements were properly classified to plasmid bin using the PlasFlow approach. We also

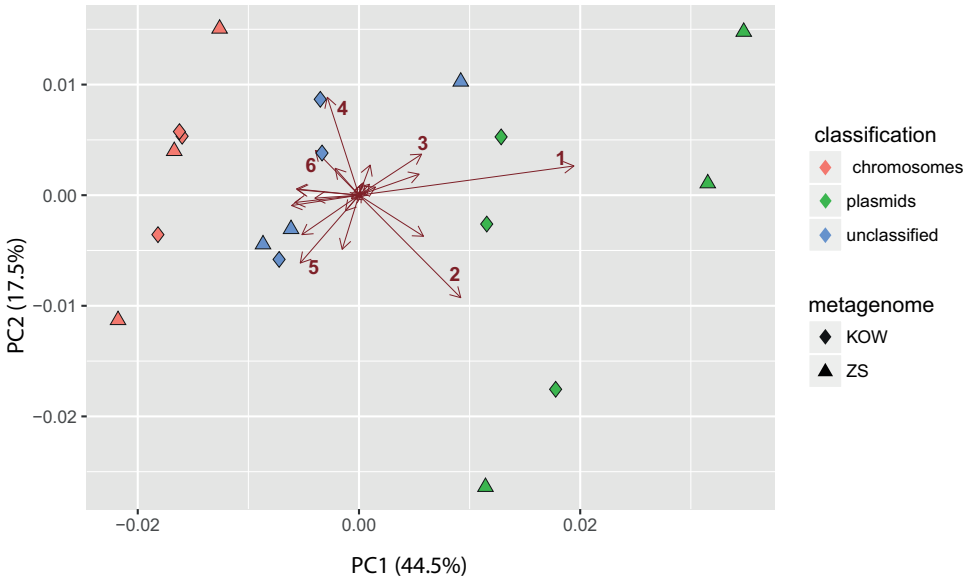

Figure 5. Principal Component Analysis (PCA) plot showing the variation between the PlasFlow-predicted plasmid, chromosomal, and unclassified sequence bins in terms of level1 subsystems abundance. Samples from uranium (KOW) and gold (ZS) mine microbial mats were assembled using SPAdes 3.9.1 and classified using PlasFlow. Obtained sequence bins were annotated using MG-RAST and analyzed using PCA to identify functions mostly contributing to the variance between classification bins. Individual bins. Colored symbols correspond to individual classification bins for each sample, where color code for a type of classification and shape for the source of sample. Vectors indicate the direction and strength of each subsystem to the overall distribution. (1) Phages, prophages, transposable elements, plasmids, (2) DNA metabolism, (3) membrane transport, (4) amino acids and derivatives, (5) respiration, (6) carbohydrates.

found that sequences classified as chromosomes are richer in functions related to basal metabolism.

Although there is a clear separation of plasmid-classified sequences from chromosomal ones, there are only minor differences between plasmid samples from different environments. Functions directly related to heavy metal resistance were represented at similar levels in both plasmidomes, with cobalt-zinc-cadmium, mercury and arsenic resistance being the most abundant and most discriminated from the chromosomal sequences (Supplemental Figure S3, Table S10).

When the PlasFlow predictions were compared to those from cBar, it revealed that only 42.63% of predictions are shared and only 35% of them are for plasmids (Supplemental Table S9), which is in contrast to the reference plasmidome datasets. cBar identified 29.91–37.28% of sequences to be of plasmid origin, which constituted 21.66–33.55% of cumulative scaffold length and more than that identified by PlasFlow (Supplemental Table S9). Annotations of corresponding bins from PlasFlow and cBar were subjected to PCA, which indicated that both are able to functionally separate plasmids from chromosomes (Supplemental Figure S2). However, differential analysis performed with DESeq2 showed that PlasFlow allowed for higher enrichment of plasmid-related functions (mostly the Phages, Prophages, Transposable Elements, and Plasmids subsystem) in the plasmid bin compared to the chromosomal bin (Supplemental Figure S4, Table S10). Nevertheless, without additional experimental data, it is impossible to say which classification is better as the actual content of plasmids in the analyzed samples is unknown.

When PlasmidFinder was applied to our metagenomic dataset, we found that this software is not suitable for analysis of such complex data as it recovered single contigs in only two out of six analyzed metagenomic samples (Supplemental Table S9). Although we expected a low performance of PlasmidFinder due to a high fragmentation of metagenomic assemblies, this result was surprising for us since we expected high representation of plasmids based on PlasFlow and hence more hits from PlasmidFinder. On the other hand, Recycler identified 35–203 possible plasmid sequences, some of which were as long as 394.1 kb. However, their annotations contained mostly phage-related genes (on average 30.3% annotations for KOW and 36.5% for ZS came from the Phages or Prophages subsystem, Supplemental Table S10) and, unlike PlasFlow, they were not attributed to integrases, but rather to r1 streptococcal phage protein. This finding suggests that phage genomes were assembled together with plasmids, which is not surprising considering that Recycler relies on the circularity of sequences and therefore cannot differentiate between plasmids and circular phages. This feature of Recycler was also reported by Arredondo-Alonso *et al.* (31).

As a result of our analyses, we showed that PlasFlow can be successfully used for the prediction of plasmid sequences in genomic and metagenomic datasets. Detailed analyses showed that it presents superior performance over other available tools in terms of accuracy and can be easily adopted for comparative plasmidome analyses.

DISCUSSION

In the present work, we have shown that simple sequence content statistics based on the kmer frequency can be used

to determine if a given sequence originates from the plasmid or chromosome. Although similar approaches exist, aimed mainly at taxonomic classification of metagenomic shotgun sequences (47,63–64), there is only one known tool (cBar) that exploits genome signatures for plasmid prediction (30). We propose a novel approach, PlasFlow, which uses a neural network-based model for the classification of metagenomic sequences and showed that it clearly outperforms cBar and other tools that use data other than genome signatures to recover plasmid sequences from sequencing data.

Algorithm rationale and comparison to other approaches

Newly developed PlasFlow, which does not rely on any prior assumptions about the taxonomical or functional composition of analyzed samples, may be successfully applied to the prediction of plasmids in genomic and metagenomic assemblies. Comparison of PlasFlow to other tools aimed at plasmid identification revealed that PlasFlow has higher accuracy and is better suited for metagenome assemblies.

Although the rationale behind the PlasFlow algorithm is simple and similar to that used by cBar (30), PlasFlow is better suited for accurate plasmid prediction of metagenomic sequences. Unlike cBar that uses self-organizing maps (SOMs), PlasFlow implements a deep artificial neural network, which is better suited for finding hidden structures in highly complicated biological data (50). In the case of cBar, training was performed using genome signatures of full-length sequences, whereas, in PlasFlow, we trained a network with the signatures from shorter sequence fragments, which makes the classifier better suited for highly fragmented metagenomic assemblies. We chose to train our neural networks using fragments of 10 kb length, since we hypothesized that their length is similar to the average contig length of typical metagenomic assembly (55) and they are long enough to assure proper kmer coverage, even for hexamers or heptamers. We divided the training dataset by the mean of the taxonomic origin (phylum), significantly improving PlasFlow performance over cBar results and allowing for more precise classification, as it eliminates a bias originating from inter-phyla differences in oligonucleotides usage (Figure 2). We also introduced probability filtering that excluded uncertain predictions; thus, the number of false hits was significantly lower, allowing for more reliable plasmidome analyses. Finally, comparison of both programs revealed that PlasFlow outperformed cBar in plasmid identification, specifically when using testing datasets comprising of different length fragments of chromosomes and plasmids downloaded from the RefSeq database (Figure 3B; Supplemental Table S5) and a set of 42 plasmid-rich bacterial genomes (Supplemental Table S2).

Although we have shown that the %G+C content differs significantly between phyla, as well as between plasmids and chromosomes from the same taxonomic group, the observed classification accuracy cannot be simply explained by differences in the %G+C content. The %G+C content simplifies nucleotide composition down to a single parameter with known limitations for investigating genome dynamics, whereas oligonucleotide frequencies capture the species-specific characteristics of nucleotide composition more effectively than %G+C (65). Therefore, the approach implemented by PlasFlow, using a wide spectrum of oligonucleotide frequencies, is advantageous over classical %G+C content analyses and can recover subtle differences between analyzed sequences.

Importantly, PlasFlow can be applied to any assembly data, even when the raw sequencing data are absent. This is in contrast to algorithms implemented in PlasmidSpades or Recycler, which require an assembly graph and paired sequencing reads to recover plasmids from the assembly. Direct comparison performed on the set of bacterial genomes and plasmids (31) revealed that PlasFlow outperforms other tools in the rate of plasmid identification. Strikingly, in the case of metagenomic data, tools that rely on information other than sequence composition performed much worse than PlasFlow, with the most prominent example being PlasmidFinder, which identified only two plasmids across six analyzed metagenomes (Supplemental Table S9).

Limitations

Classification of assembled sequences by PlasFlow can be disturbed by several factors. First, if a sequence has a signature that differs from what was included in the training set, the system will produce small probability values for all possible classes and the sequence will not be properly classified. This is a common limitation of supervised machine learning approaches, as it is impossible to recognize something that was not seen before. Therefore, since the training set is limited to the actual content of sequence databases, some sequences will not be properly classified. Another risk is the existence of chromosomes with signatures very close to those of plasmids used for training, which will be improperly classified as plasmids (false positives). This can only be overcome by regularly updating the model with newly published genomic data. Second, an assembly itself can be erroneous, especially when applied to phylogenetically diverse samples when the risk of creating interspecies chimeras is very high (66). In the context of plasmid sequences, the risk is even higher due to similar backbone elements, such as replication and conjugation transfer elements (8,67) and a large number of mobile genetic elements, such as insertion sequences and transposons (68), which are similarly structured. Third, plasmid sequences can sometimes be integrated into genomes, making differentiation between chromosomes and plasmids more difficult. Fourth, it is more difficult to obtain a proper sequence signature for short sequences, as many kmers cannot be covered. However, short sequences are generally less informative since they carry a small number of genes and can usually be isolated by standard molecular biology techniques; for most of the known metagenomic DNA isolation methods fragments of 1–10 kb remain intact. Hence, short sequences should be avoided in the PlasFlow-based classification as they may be problematic for the classifier and lead to erroneous predictions. Fifth, it is usual for oligonucleotide-based classification procedures that 100% accuracy cannot be achieved, therefore the user should expect that some of predictions will be wrong, especially for short sequences. Nevertheless, PlasFlow has extraordinary low false predictions rate (close

to 5%) which is remarkable, especially in comparison to cBar for which erroneous predictions are four times more frequent (Supplemental Table S5). Taking all above into account, excluding low-probability classifications is advantageous for PlasFlow and increases overall accuracy.

In contrast to tools like PlasmidSpades or Recycler, which output possible full-length plasmid sequences based on their circularity or differential sequencing coverage, PlasFlow (similarly to cBar) can predict if a given sequence might be of plasmid origin or not, even if it does not cover the entire plasmid sequence. Therefore, PlasFlow can be useful in the gene-centric approaches where there is no need to obtain the full plasmid sequence (such as with most of metagenomic analyses, which we have shown using mine samples), but it cannot be used in other cases where the full sequence is desired as well as the precise taxonomic information. However, classification obtained from PlasFlow can provide a nice starting point for more detailed experiments and analyses.

Applicability

PlasFlow is a tool developed to identify plasmid sequences in genomic and metagenomic assemblies and is especially useful for analysis of large plasmids and megaplasmids, which are often missed in most of the currently used standard molecular biology techniques (19) and therefore may be a valuable extension of plasmidome studies. Unlike currently used protocols for the plasmidome research, PlasFlow does not rely on the circularity of assembled sequences as in the case of standard plasmidome isolation techniques (19) or the Recycler algorithm (24), allowing for the description of plasmidomes, even if only a low quality (fragmented) assembly is present without the careful isolation of single plasmids. This approach also extends the spectrum of analyzed plasmidome sequences to linear plasmids, which have been vastly ignored in previous plasmidome studies (3). Additionally, due to the taxon-specific genome signatures used for training, PlasFlow can predict phyla of analyzed sequences, providing information about taxons that mostly contribute to the plasmidome in a given environment.

The ecological impact of plasmids is significant, as they commonly carry genes that foster diversification and adaptation of bacterial populations (1). We have analyzed the role of plasmids in a heavy-metal contaminated environments from abandoned mines and PlasFlow was able to identify plasmid sequences carrying genes involved in heavy-metal resistance, much less represented in the chromosomal sequences.

Although plasmid-encoded functions are beneficial for bacteria, they can cause disturbances to human health with regard to virulence factors and the antibiotic resistance, which can easily spread between bacterial species through the horizontal gene transfer (4,69–70). Therefore plasmidome studies are important for our understanding of the evolution of bacterial communities. We believe that advantages of PlasFlow will make it a standard tool included in the metagenomic and plasmidome analysis pipelines.

DATA AVAILABILITY

PlasFlow is an open source software available at GitHub (https://github.com/smaegol/PlasFlow). PlasFlow-classified contigs from the KOW and ZS assemblies have been deposited with the MG-RAST (http://metagenomics.anl.gov) under the project with the accession number mgp80771.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge Marta Stepniewska-Dziubinska and Maciej Wojcikowski for help with machine learning, and Dorota Adamska, Sylwia Czarnomska, Mariusz Czarnocki-Cieciura, Zbigniew Pietras and Adam Pyzik for critical reading of this manuscript. We sincerely thank anonymous reviewers whose comments substantially helped to improve this manuscript.

FUNDING

National Science Centre [Preludium, UMO-2012/05/N/NZ9/01393 to P.S.K.]; EU European Regional Development Fund; Operational Program Innovative Economy 2007–2013 [POIG.01.01.02-14-054/09-00]; European Social Fund, Human Capital Operational Program for the execution of the project 'Support for bio tech med scientists in technology transfer' [UDA-POKL.08.02.01-14-041/09], to P.K.; Polish Ministry of Science and Higher Education [POIG.02.02.00-14-024/08-00 and POIG.02.03.00-00-003/09-00]. Funding for open access charge: IBB PAS Statutory Funds.
Conflict of interest statement. None declared.

REFERENCES

1. Heuer,H. and Smalla,K. (2012) Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol. Rev.*, **36**, 1083–1104.
2. Heuer,H., Abdo,Z. and Smalla,K. (2008) Patchy distribution of flexible genetic elements in bacterial populations mediates robustness to environmental uncertainty. *FEMS Microbiol. Ecol.*, **65**, 361–371.
3. Dib,J.R., Wagenknecht,M., Farías,M.E. and Meinhardt,F. (2015) Strategies and approaches in plasmidome studies—uncovering plasmid diversity disregarding of linear elements? *Front. Microbiol.*, **6**, 463.
4. Carattoli,A. (2013) Plasmids and the spread of resistance. *Int. J. Med. Microbiol.*, **303**, 298–304.
5. Segura,A., Molina,L. and Ramos,J.L. (2014) Plasmid-mediated tolerance toward environmental pollutants. *Microbiol. Spectr.*, **2**, PLAS-0013-2013.
6. Zhang,T., Zhang,X.-X. and Ye,L. (2011) Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PLoS ONE*, **6**, e26041.
7. Thomas,C.M. and Nielsen,K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.*, **3**, 711–721.
8. Shintani,M., Sanchez,Z.K. and Kimbara,K. (2015) Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Evol. Genomic Microbiol.*, **6**, 242.

9. Bleicher,A., Schöfl,G., Rodicio,M.D.R. and Saluz,H.P. (2013) The plasmidome of a Salmonella enterica serovar Derby isolated from pork meat. *Plasmid*, **69**, 202–210.
10. Brolund,A., Franzén,O., Melefors,Ö., Tegmark-Wisell,K. and Sandegren,L. (2013) Plasmidome-analysis of ESBL-producing Escherichia coli using conventional typing and high-throughput sequencing. *PLoS ONE*, **8**, e65793.
11. Fondi,M., Bacci,G., Brilli,M., Papaleo,M.C., Mengoni,A., Vaneechoutte,M., Dijkshoorn,L. and Fani,R. (2010) Exploring the evolutionary dynamics of plasmids: the Acinetobacter pan-plasmidome. *BMC Evol. Biol.*, **10**, 59.
12. Szczepanowski,R., Bekel,T., Goesmann,A., Krause,L., Krömeke,H., Kaiser,O., Eichler,W., Pühler,A. and Schlüter,A. (2008) Insight into the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to antimicrobial drugs analysed by the 454-pyrosequencing technology. *J. Biotechnol.*, **136**, 54–64.
13. Bale,M.J., Day,M.J. and Fry,J.C. (1988) Novel method for studying plasmid transfer in undisturbed river epilithon. *Appl. Environ. Microbiol.*, **54**, 2756–2758.
14. Hill,K.E., Weightman,A.J. and Fry,J.C. (1992) Isolation and screening of plasmids from the epilithon which mobilize recombinant plasmid pD10. *Appl. Environ. Microbiol.*, **58**, 1292–1300.
15. Jones,B.V. and Marchesi,J.R. (2007) Accessing the mobile metagenome of the human gut microbiota. *Mol. Biosyst.*, **3**, 749–758.
16. Jones,B.V. and Marchesi,J.R. (2007) Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods*, **4**, 55–61.
17. Jones,B.V., Sun,F. and Marchesi,J.R. (2010) Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. *BMC Genomics*, **11**, 46.
18. Brown Kav,A., Sasson,G., Jami,E., Doron-Faigenboim,A., Benhar,I. and Mizrahi,I. (2012) Insights into the bovine rumen plasmidome. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5452–5457.
19. Brown Kav,A., Benhar,I. and Mizrahi,I. (2013) A method for purifying high quality and high yield plasmid DNA for metagenomic and deep sequencing approaches. *J. Microbiol. Methods*, **95**, 272–279.
20. Li,A.-D., Li,L.-G. and Zhang,T. (2015) Exploring antibiotic resistance genes and metal resistance genes in plasmid metagenomes from wastewater treatment plants. *Front. Microbiol.*, **6**, 1025.
21. Jørgensen,T.S., Xu,Z., Hansen,M.A., Sørensen,S.J. and Hansen,L.H. (2014) Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS ONE*, **9**, e87924.
22. Kristiansson,E., Fick,J., Janzon,A., Grabic,R., Rutgersson,C., Weijdegård,B., Söderström,H. and Larsson,D.G.J. (2011) Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS ONE*, **6**, e17038.
23. Antipov,D., Hartwick,N., Shen,M., Raiko,M., Lapidus,A. and Pevzner,P.A. (2016) plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, **32**, 3380–3387.
24. Rozov,R., Brown Kav,A., Bogumil,D., Shterzer,N., Halperin,E., Mizrahi,I. and Shamir,R. (2017) Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*, **33**, 475–482.
25. Carattoli,A., Zankari,E., García-Fernández,A., Larsen,M.V., Lund,O., Villa,L., Aarestrup,F.M. and Hasman,H. (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.
26. Lanza,V.F., de Toro,M., Garcillán-Barcia,M.P., Mora,A., Blanco,J., Coque,T.M. and de la Cruz,F. (2014) Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet.*, **10**, e1004766.
27. Bohlin,J., Skjerve,E. and Ussery,D.W. (2008) Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics*, **9**, 104.
28. Bohlin,J., van Passel,M.W., Snipen,L., Kristoffersen,A.B., Ussery,D. and Hardy,S.P. (2012) Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. *BMC Genomics*, **13**, 66.
29. Zhou,F., Olman,V. and Xu,Y. (2008) Barcodes for genomes and applications. *BMC Bioinformatics*, **9**, 546.
30. Zhou,F. and Xu,Y. (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, **26**, 2051–2052.
31. Arredondo-Alonso,S., Willems,R.J., van Schaik,W. and Schürch,A.C. (2017) On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics*, **3**, e000128
32. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
33. Sentchilo,V., Mayer,A.P., Guy,L., Miyazaki,R., Green Tringe,S., Barry,K., Malfatti,S., Goessmann,A., Robinson-Rechavi,M. and van der Meer,J.R. (2013) Community-wide plasmid gene mobilization and selection. *ISME J.*, **7**, 1173–1186.
34. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
35. Uhrynowski,W., Decewicz,P., Dziewit,L., Radlinska,M., Krawczyk,P.S., Lipinski,L., Adamska,D. and Drewniak,L. (2017) Analysis of the genome and mobilome of a dissimilatory arsenate reducing Aeromonas sp. O23A reveals multiple mechanisms for heavy metal resistance and metabolism. *Front. Microbiol.*, **8**, 936.
36. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
37. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
38. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
39. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
40. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
41. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
42. Saeed,I., Tang,S.-L. and Halgamuge,S.K. (2012) Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.*, **40**, e34–e34.
43. Nishida,H. (2012) Comparative analyses of base compositions, DNA sizes, and Dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. *Int. J. Evol. Biol.*, **2012**, e342482.
44. Suzuki,H., Sota,M., Brown,C.J. and Top,E.M. (2008) Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res.*, **36**, e147–e147.
45. van Passel,M.W., Bart,A., Luyf,A.C., van Kampen,A.H. and van der Ende,A. (2006) Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics*, **7**, 26.
46. Abe,T., Kanaya,S., Kinouchi,M., Ichiba,Y., Kozuki,T. and Ikemura,T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res.*, **13**, 693–702.
47. McHardy,A.C., Martín,H.G., Tsirigos,A., Hugenholtz,P. and Rigoutsos,I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.
48. Patil,K.R., Haider,P., Pope,P.B., Turnbaugh,P.J., Morrison,M., Scheffer,T. and McHardy,A.C. (2011) Taxonomic metagenome sequence assignment with structured output models. *Nat. Methods*, **8**, 191–192.
49. Willner,D., Thurber,R.V. and Rohwer,F. (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.*, **11**, 1752–1766.
50. Angermueller,C., Pärnamaa,T., Parts,L. and Stegle,O. (2016) Deep learning for computational biology. *Mol. Syst. Biol.*, **12**, 878.
51. Lee,S., Kong,S. and Xing,E.P. (2016) A network-driven approach for genome-wide association mapping. *Bioinformatics*, **32**, i164–i173.

52. Rampasek,L. and Goldenberg,A. (2016) TensorFlow: biology's gateway to deep learning? *Cell Syst.*, **2**, 12–14.
53. Jha,A., Gazzara,M.R. and Barash,Y. (2017) Integrative deep models for alternative splicing. *Bioinformatics*, **33**, i274–i282.
54. Bohlin,J., Skjerve,E. and Ussery,D.W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLOS Comput. Biol.*, **4**, e1000057.
55. Vollmers,J., Wiegand,S. and Kaster,A.-K. (2017) Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS ONE*, **12**, e0169662.
56. Dietterich,T.G. (2000) Ensemble methods in machine learning. In: *Multiple Classifier Systems*. Springer, Berlin, Heidelberg, pp. 1–15.
57. Anda,M., Ohtsubo,Y., Okubo,T., Sugawara,M., Nagata,Y., Tsuda,M., Minamisawa,K. and Mitsui,H. (2015) Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 14343–14347.
58. Battermann,A., Disse-Krömker,C. and Dreiseikelmann,B. (2003) A functional plasmid-borne rrn operon in soil isolates belonging to the genus Paracoccus. *Microbiology*, **149**, 3587–3593.
59. Kunnimalaiyaan,M., Stevenson,D.M., Zhou,Y. and Vary,P.S. (2001) Analysis of the replicon region and identification of an rRNA operon on pBM400 of Bacillus megaterium QM B1551. *Mol. Microbiol.*, **39**, 1010–1021.
60. Drewniak,L., Krawczyk,P.S., Mielnicki,S., Adamska,D., Sobczak,A., Lipinski,L., Burec-Drewniak,W. and Sklodowska,A. (2016) Physiological and metagenomic analyses of microbial mats involved in self-purification of mine waters contaminated with heavy metals. *Front. Microbiol.*, **7**, 1252.
61. Bellanger,X., Payot,S., Leblond-Bourget,N. and Guédon,G. (2014) Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.*, **38**, 720–760.
62. Unterholzner,S.J., Poppenberger,B. and Rozhon,W. (2013) Toxin–antitoxin systems: biology, identification, and application. *Mob. Genet. Elem.*, **3**, e26219.
63. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
64. Hou,T., Liu,F., Liu,Y., Zou,Q.Y., Zhang,X. and Wang,K. (2015) Classification of metagenomics data at lower taxonomic level using a robust supervised classifier. *Evol. Bioinforma.*, **11**, 3–10.
65. Dick,G.J., Andersson,A.F., Baker,B.J., Simmons,S.L., Thomas,B.C., Yelton,A.P. and Banfield,J.F. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol.*, **10**, R85.
66. Wooley,J.C., Godzik,A. and Friedberg,I. (2010) A primer on metagenomics. *PLOS Comput. Biol.*, **6**, e1000667.
67. Smillie,C., Garcillán-Barcia,M.P., Francia,M.V., Rocha,E.P.C. and de la Cruz,F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.
68. Siguier,P., Filée,J. and Chandler,M. (2006) Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.*, **9**, 526–531.
69. Djordjevic,S.P., Stokes,H.W. and Roy Chowdhury,P. (2013) Mobile elements, zoonotic pathogens and commensal bacteria: conduits for the delivery of resistance genes into humans, production animals and soil microbiota. *Front. Microbiol.*, **4**, 86.
70. Singer,A.C., Shaw,H., Rhodes,V. and Hart,A. (2016) Review of antimicrobial resistance in the environment and its relevance to environmental regulators. *Front. Microbiol.*, **7**, 1728.